Predictors of Trust in AI

Summary of Preliminary Results

Scott Claessens

2025-09-24

Table of contents

This document summarises the preliminary results of our study on the predictors of trust in artificial intelligence systems.

We recruited 800 participants from the UK (N = 690 after exclusions for low Captcha scores and failed attention check) from Prolific. In the study, participants answered questions about both “general AI” and a random subset of specific AI systems (5 out of 20). In each case, we asked participants how much they trusted the system and various other questions, including perceived reliability, competence, genuine nature, ethicality, autonomy, potential for good, potential for harm, interpretability, explainability, anthropomorphism, and predictability. We also collected data on whether participants had heard of and used the different AI systems.

Data and code for the study can be found here: <https://github.com/ScottClaessens/person-trust-AI>

## Reported usage of different AI systems

We found that some AI systems were more familiar to participants than others.

|  |
| --- |
| Figure 1: Proportion of participants that have heard of and used different AI systems |

## Rankings for different AI systems

We ranked average ratings of trust for “general AI” and the different AI systems. The results show that participants trust some AI systems more than others, with “general AI” sitting somewhere in the middle.

|  |
| --- |
| Figure 2: Self-reported trust in different AI systems |

Rankings for other variables are below:

|  |
| --- |
| Figure 3: Perceived reliability of different AI systems |

|  |
| --- |
| Figure 4: Perceived competence of different AI systems |

|  |
| --- |
| Figure 5: Perceived genuine nature of different AI systems |

|  |
| --- |
| Figure 6: Perceived ethicality of different AI systems |

|  |
| --- |
| Figure 7: Perceived autonomy of different AI systems |

|  |
| --- |
| Figure 8: Perceived ‘potential good’ of different AI systems |

|  |
| --- |
| Figure 9: Perceived ‘potential harm’ of different AI systems |

|  |
| --- |
| Figure 10: Perceived interpretability of different AI systems |

|  |
| --- |
| Figure 11: Perceived explainability of different AI systems |

|  |
| --- |
| Figure 12: Perceived anthropomorphism of different AI systems |

|  |
| --- |
| Figure 13: Perceived predictability of different AI systems |

## Predictors of trust in general AI

For general AI, we looked at the correlation matrix for self-reported trust and all other questions.

|  |
| --- |
| Figure 14: Correlation matrix for ‘General AI’ |

We found that perceived reliability, competence, genuine nature, and ethicality have the strongest positive correlations with trust. This makes sense, since previous work has suggested that trust in AI is a multidimensional concept that is made up of both performance trust (reliability and competence) and moral trust (genuine nature and ethicality; Malle & Ullman, 2021). Trust is also strongly positively correlated with the belief that AI has the potential for good.

Other variables are also correlated with trust, but to a lesser degree.

## Predictors of trust in different AI systems

To assess how this pattern varies across different AI systems, we fitted a series of multilevel models that allowed the effects of predictors to vary across AI systems. The variance parameter for the varying slopes allows us to determine how much the effect of each predictor on trust *varies* across different AI systems.

|  |
| --- |
| Figure 15: Variance in slopes across different AI systems |

This plot suggests that, for example, the effect of interpretability on trust does not vary much across different AI systems. On the other hand, the effect of competence on trust varies much more across different AI systems. These effects are shown in the plots below.

Plots for all other predictors of trust are displayed below:

|  |
| --- |
| Figure 16: Effect of reliability on trust across different AI systems |

|  |
| --- |
| Figure 17: Effect of competence on trust across different AI systems |

|  |
| --- |
| Figure 18: Effect of ‘genuine’ on trust across different AI systems |

|  |
| --- |
| Figure 19: Effect of ethicality on trust across different AI systems |

|  |
| --- |
| Figure 20: Effect of autonomy on trust across different AI systems |

|  |
| --- |
| Figure 21: Effect of potential good on trust across different AI systems |

|  |
| --- |
| Figure 22: Effect of potential harm on trust across different AI systems |

|  |
| --- |
| Figure 23: Effect of interpretability on trust across different AI systems |

|  |
| --- |
| Figure 24: Effect of explainability on trust across different AI systems |

|  |
| --- |
| Figure 25: Effect of anthropomorphism on trust across different AI systems |

|  |
| --- |
| Figure 26: Effect of predictability on trust across different AI systems |

## Future directions

There is more that we can do with this dataset:

* **Which types of AI are most prototypical?** We pre-registered that we would conduct a principal components analysis on all the variables and map the different AI types in the dimensional space to determine which are closest to “general AI”. In other words, which types of AI are people probably thinking about when surveys ask them about AI in general? Unfortunately for us, everything is highly correlated in this dataset, meaning that PCA just pulls out a single dimension (which can be conceptualised as how “positive” or “negative” participants feel).
* **How does trust vary across individuals?** We have data on demographics, political ideology, and familiarity with AI, so we could potentially use these as predictors of trust.
* **Are there any interaction effects?** We have looked at predictors separately. But the variables likely interact with each other. For example, competence is probably a weaker predictor of trust for autonomous killer drones because participants believe that this AI has the potential for harm – they don’t want it to achieve that harm more competently!

Beyond these questions, we would like to scale this up across countries. Before doing this, it would be good to get some direction on what is most interesting to focus on going forward.

## References

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction* (pp. 3-25). Academic Press.